Managing Data Exercise 2

The following is the SQL script for generating a repeatable random sample as discussed in class (There are no functional problems with this SQL script, but there is a "not so pretty problem." If you can figure out what that not so pretty problem is, email me (bruce) before 11:59 pm ET on October 2nd. I will reward you with 0.5 points toward your final course grade (it may help in a borderline situation 😊). Please do not share with others your answer to this "not so pretty" problem before 11:59 pm ET. We will post a thread on Ed Discussion for you to share after October 2nd.

The output of this SQL script will be stored in a CSV file:

```
SELECT 'Temperature', 'Sales' FROM cookies.sales

UNION

(SELECT DISTINCT Temperature, Sales FROM cookies.sales

ORDER BY RAND(7)

LIMIT 25)

INTO OUTFILE 'C:/ProgramData/MySQL/MySQL Server 8.0/Uploads/test100.csv'

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n';
```

Use the nj_state_teachers_salaries database (the cleaned version – by you) to complete the following tasks:

1. Write a Python program to clean the data using the technique discussed in class (eliminate all blank lines, change all invalid values to NaN (np.NaN), drop all rows with blank values (NaN) in any columns. Save your cleaned data back in the CSV file. Create your database (nj_state_teachers_salaries) and table (nj_state_teachers_salaries). Load the table with your cleaned version of the data using the Python MySQL Connector.

2. Using MySQL workbench or command line, write a SQL script to perform the following tasks. In your submitted file, you should assume the existence of the new_nj_state_teachers_salaries table in the nj_state_teachers_salaries database.:
    a. Use a SELECT statement to generate and output a random sample to :
        - Include all columns
        - Include field (column) headings
        - Randomly select 777 records with a seed value of 7
        - Output results to a CSV file named teachersample.csv
    b. Create a new database called teacher_sample and a table named teachers using teachersample.csv
    c. Using the teacher_sample database, perform the following tasks:
        - Calculate the average salary

- Calculate the number of people whose salary is more than 150,000.
- Get the last name of the ones who make more than 150,000 but has less than 5 years of total experience
- Get the highest salary for Preschool, School Counselor, Principal (anyone with the word Principal in the title) , School Psychologist, and Kindergarten.
- Get the last name, first name, and salary of the lowest earner who works in Atlantic City

d. Get the total number of employees working in Passaic City with more than ten years of total experience.

What you need to submit:

1. Your Python program (name it nj_clearner.py)
2. your SQL script (name it nj_teachers.sql)